SALIENCE CYBER | THREAT RESEARCH
White Paper

Defending the AI-Enabled Workplace

A Representative-Sample Analysis of Browser- and AI-Workflow Threat Trends and the Mitigation Efficacy of the Salience Cyber User Edge Platform

Salience Cyber Threat Research · May 2026

ABSTRACT

The integration of artificial intelligence into routine business workflows is now effectively non-optional: organizations that defer adoption forfeit competitive position, while those that adopt extend their attack surface into channels that conventional security tooling was not designed to observe. This paper presents a representative-sample analysis of threat activity recorded by the Salience Cyber User Edge platform across a nine-day operational window. The dataset — 311 discrete threat events captured by five sensors — is offered not as a population-level census but as a directional sample sufficient to characterize the relative prevalence of threat categories and to demonstrate platform mitigation behavior under live conditions. Established threats (command-and-control traffic, malicious-domain connections, and script injection) dominate by volume, while a small but qualitatively significant set of events — prompt injection and data exfiltration targeting AI application interfaces — signal an emerging class of AI-workflow threats largely invisible to perimeter and signature-based controls. Within the sample, 92.6% of events were mitigated in real time without user interruption. We discuss the implications for securing AI-enabled work, and the limitations inherent to single-window operational telemetry.

Keywords: *browser security; AI-workflow security; prompt injection; behavioral detection; threat telemetry; representative sampling; command-and-control.*

1. Introduction

Artificial intelligence has moved, within a remarkably short period, from a discretionary technology to an operational expectation. Large language models now draft correspondence, generate and review code, summarize documents, and inform decisions across nearly every business function. For most organizations the relevant question is no longer whether to adopt AI, but how to do so without incurring unacceptable risk. AI adoption is best treated not as a trend to be evaluated but as a condition to be secured around.

That condition has a direct security consequence. The locus of knowledge work has shifted away from the defensible network perimeter — the boundary that firewalls, network appliances, and traditional endpoint tools were architected to protect — and into the browser, into software-as-a-service applications, and into AI-mediated workflows through which an organization's most sensitive material now routinely passes. Threat activity has followed that shift. Adversary tooling that was once the preserve of well-resourced state actors is now widely available; malicious code is engineered specifically to defeat signature-based detection; and a new category of attack targets AI systems directly, through inputs crafted to manipulate model behavior rather than to exploit a software flaw.

This paper examines that landscape empirically. Drawing on operational telemetry from the Salience Cyber User Edge platform, it characterizes the threat categories users encounter in practice, identifies directional trends — including the emergence of AI-workflow threats — and reports the platform's mitigation outcomes under live conditions. The analysis is deliberately framed around a representative

sample rather than a comprehensive census; Section 2 sets out what that framing does and does not support.

2. Data and Methodology

2.1 Source and scope

The dataset analyzed in this paper comprises 311 discrete threat events recorded by the Salience Cyber User Edge platform across five active sensors during a continuous nine-day operational window (23 March – 1 April 2026). Each event was logged to a central console at the time of detection, together with its category, an automatically assigned severity classification, the detection layer responsible, and the mitigation outcome. An event is defined here as a single discrete detection; aggregate figures throughout this paper are counts of such events.

2.2 Representative sampling: what the data supports

This paper characterizes its dataset explicitly and deliberately as a representative sample rather than a population-level census. The observation window is short, the sensor count is modest, and the deployment context is specific; the data therefore should not be read as an estimate of absolute threat incidence across all environments, nor as a basis for statistical generalization with formal confidence bounds.

What a sample of this nature does reliably support is threefold. *First*, it permits directional characterization of the *relative* prevalence of threat categories — which classes of threat are common, and which are rare in practice. *Second*, it provides direct observational evidence that threat types are active in the wild, which is sufficient to establish their existence and qualitative character regardless of sample size. *Third*, it demonstrates the mitigation behavior of the platform under genuine operational conditions, since every recorded event carries a corresponding outcome. Conclusions in this paper are confined to these three classes of claim.

2.3 Classification

Each event was assigned a confidence score between 0.0 and 1.0 by the platform, which in turn drove an automated severity classification applied consistently across all sensors, sessions, and threat types. Severity classes and category labels are those of the platform's own taxonomy; they are internally consistent and auditable, but, as noted in Section 7, they are not externally calibrated against an independent standard.

3. Threat Landscape Trends

Three structural trends are evident in the sample, alongside one finding that is small in volume but disproportionate in significance.

3.1 Established threats dominate by volume

Most of the recorded activity belongs to well-documented threat classes. Command-and-control (C2) framework traffic alone accounts for 38.9% of events, followed by malicious-domain connections (18.3%), Content-Security-Policy violations (12.2%), and cross-site scripting (9.0%). Collectively, threats amenable in principle to signature- or rule-based recognition constitute the bulk of the sample. The implication is not that legacy detection is obsolete, but that it remains necessary while being, on the evidence below, insufficient.

3.2 The commoditization of advanced tooling

C2 callbacks within the sample originated from seventeen distinct tool families, including frameworks — Cobalt Strike, Sliver, Brute Ratel, Havoc, and Mythic among them — that are documented components of nation-state and major ransomware operations [1], [2], [3]. The breadth of tooling observed over only nine days indicates that adversary capability once confined to well-resourced actors is now broadly accessible. This is consistent with the wider literature on extended attacker dwell time, during which such tooling is used to establish persistence, harvest credentials, and stage data prior to discovery [2], [4].

3.3 The abuse of trusted infrastructure

Script-injection and polymorphic-malware activity in the sample was observed predominantly on high-reputation, high-traffic destinations rather than on obviously suspicious domains. This pattern reflects a deliberate adversary strategy: trusted destinations attract implicit trust from users and security tooling alike and serve as effective delivery infrastructure. It also has a clear architectural consequence — reputation-based and perimeter-oriented models, which assume that risk correlates with the apparent trustworthiness of a destination, are structurally ill-suited to this threat pattern.

3.4 The emergence of AI-workflow threats

Four events in the sample — two prompt-injection attempts and two data-exfiltration events, all associated with AI application interfaces — are small in count but represent the most consequential finding of this analysis. Prompt injection exploits no conventional software vulnerability; it manipulates model behavior through adversarial input and leaves no network signature for a firewall to match, no file for an endpoint agent to scan, and no event pattern that a conventional alert was written to detect. Data exfiltration to a legitimate AI interface is similarly difficult to distinguish, at the network layer, from sanctioned use. These events are direct observational evidence that AI workflows are already an active attack surface, not a prospective one.

4. Threat Types Encountered

Table 1 sets out the full distribution of threat categories within the sample, with representative vectors for each. The table is ordered by event volume.

Table 1. Distribution of threat categories within the nine-day sample (n = 311).

Threat Category	Events	Share	Peak Severity	Representative Vectors
Command & Control framework	121	38.9%	Critical	Implant callbacks across 17 distinct tool families (e.g., Cobalt Strike, Sliver, Brute Ratel, Havoc).
Malicious domain connection	57	18.3%	Critical	Phishing, web-skimming, exploit-kit delivery, and cryptomining infrastructure.
CSP violation	38	12.2%	Medium	Content-Security-Policy breaches indicating active script-injection pressure.
Cross-site scripting (XSS)	28	9.0%	High	Script injection within authenticated sessions on high-reputation platforms.
Suspicious connection	23	7.4%	Low	Un-codified anomalous traffic surfaced by behavioral analysis.
Polymorphic malware	20	6.4%	High	Self-modifying code delivered via trusted, high-traffic destinations.
Penetration-test tooling	15	4.8%	High	Offensive-security framework traffic in monitored environments.
XSS platform	3	1.0%	High	Hosted cross-site-scripting attack infrastructure.
Data exfiltration	2	0.6%	High	Outbound transfer of sensitive data, including to AI application interfaces.
Prompt injection	2	0.6%	Critical	Adversarial inputs targeting AI application interfaces.
Unknown / unclassified	2	0.6%	Medium	Events without an established classification.

Shares are of 311 total events. Peak severity reflects the highest recorded instance within each category.

Two observations warrant emphasis. The first concerns the behavioral-detection category labelled “suspicious connection” (7.4% of the sample): these are events that matched no established signature and were instead surfaced as anomalies. Their presence indicates a persistent tail of un-codified activity that signature-based approaches cannot, by definition, detect now it occurs. The second concerns the AI-workflow categories — prompt injection and data exfiltration — which, although they together account for under 1.5% of events, are distinguished by being largely invisible to the conventional security stack and by targeting precisely the workflows organizations are now adopting most rapidly.

5. Mitigation Efficacy

Of the 311 events in the sample, 288 — a real-time mitigation rate of 92.6% — were blocked or prevented before reaching the user in a harmful state. Of these, 204 were terminated at the network layer prior to payload execution, 46 were neutralized through runtime content analysis, and 38 were contained under active monitoring with no user-facing impact. The remaining 23 events were the behaviorally surfaced anomalies described above; these were logged for forensic analysis rather than blocked, a point addressed below and in Section 7.

Severity was concentrated toward the upper end of the scale. As Table 2 shows, 57.9% of events were classified Critical and a further 21.5% High — a combined 79.4%. Within the constraints of the platform's own classification scheme, this distribution is consistent with a detection posture calibrated toward high-confidence, high-consequence events rather than high-volume, low-fidelity alerting.

Table 2. Severity distribution within the sample.

Severity Class	Events	Share of Sample
Critical	180	57.9%
High	67	21.5%
Medium	41	13.2%
Low	23	7.4%

Severity classes are platform-assigned and internally consistent; they are not externally calibrated (see Section 7).

Mitigation outcomes are best understood by detection layer. The platform employs a layered architecture in which each layer is scoped to a distinct portion of the attack surface, such that activity evading one layer may be caught by another. Table 3 reports the contribution and immediate block rate of each layer.

Table 3. Events and immediate block rate by detection layer.

Detection Layer	Events	Immediate Block Rate	Functional Role
Network request engine	198	100%	Pre-execution blocking of C2 and malicious-domain traffic at the network layer.
Content-script detectors	86	100%	Runtime page analysis for XSS, CSP violations, and polymorphic malware.
Behavioral network monitor (HTM)	23	0%*	Surfacing of anomalous, un-codified traffic for forensic logging and rule development.
AI-workflow proxy	4	100%	Detection of prompt injection and data exfiltration within AI and developer interactions.

** The behavioral monitor records a 0% immediate block rate by design: its function is to surface novel, un-codified activity for forensic logging and subsequent rule development rather than to block traffic automatically. This is a deliberate architectural choice, not a detection failure.*

Two properties of this result are relevant to the paper's wider argument. First, the AI-workflow proxy — the layer responsible for the prompt-injection and data-exfiltration detections — extends mitigation into a channel that the other layers, and conventional tooling generally, do not observe. Second, no event in the sample triggered a user-facing security prompt; mitigation occurred without interrupting the user's workflow. For securing AI-enabled work, this second property is not incidental, as Section 6 argues.

6. Discussion: Operating Safely in an AI-Inevitable World

If AI adoption is treated as a settled condition, the central security question becomes one of coverage: which parts of the AI-enabled workflow are observed by existing controls, and which are not. The sample analyzed here indicates that the AI workflow itself is, for most organizations, an unobserved channel. Prompt injection produces no artifact that signature- or perimeter-based tooling is designed to recognize, and the transfer of sensitive data to a legitimate AI interface is, at the network layer, difficult to distinguish from sanctioned activity.

This visibility gap is compounded by a property of AI-workflow data loss that the sample illustrates directly: the mechanics of inadvertent disclosure and deliberate exfiltration are functionally identical. Sensitive material — source code, internal documentation, credentials, proprietary logic — leaving a controlled environment through an unmonitored channel constitutes the same exposure whether the cause is a malicious actor or a well-intentioned employee using an AI tool as intended. Governance policies and user training are necessary responses, but they are not sufficient ones, because they depend on consistent human judgment in precisely the high-frequency, low-salience moments in which such judgment is least reliable. Technical controls that observe the AI channel directly are required to close the gap.

The efficacy results offer a constructive counterpart to this diagnosis. They indicate that protection can be extended into AI and developer workflows — and can mitigate threats within them in real time — without degrading the user experience on which AI adoption depends. This matters because security measures that impede legitimate work are, in practice, circumvented or disabled, which converts a protective control into a false assurance. The absence of user-facing interruptions across the sample suggests that the trade-off organizations often assume — between enabling AI use and securing it — is not a necessary one.

Finally, the behaviorally surfaced anomalies underscore a structural point. Because a portion of threat activity is, at any given moment, un-codified, any approach that relies exclusively on known signatures will operate with a permanent detection lag. A behavioral layer does not eliminate that lag, but it converts novel activity into a recorded, analyzable signal from which future detection rules can be derived — shortening the interval between the first appearance of a threat and the capacity to block it systematically.

7. Limitations

Several limitations bound the conclusions of this paper and should be weighed by the reader. The dataset derives from a single platform's telemetry; the same platform both detects and classifies the events it reports, and the analysis has not been validated against independent third-party measurement. The observation window is short (nine days), and the sensor count modest (five); the observed distribution of threat categories may therefore reflect the specific deployment context and sensor placement rather than a generalizable population mix. Severity classifications, while internally consistent, are platform-assigned and not calibrated against an external standard, so the reported Critical-and-High share should be interpreted within that scheme rather than as an absolute measure. The 0% immediate block rate of the behavioral layer is a deliberate design choice favoring forensic logging; the downstream efficacy of the resulting rule development is not measured here. For all these reasons, the data support directional and qualitative conclusions — the relative prevalence of threat types, the demonstrated existence of AI-workflow threats, and observed mitigation behavior — but not

population-level incidence estimates. Multi-window, multi-environment study would be required to extend the findings further.

8. Conclusion

The adoption of artificial intelligence into business workflows is, for practical purposes, a settled condition; the meaningful security question is how to operate safely within it rather than whether to permit it. The representative sample examined in this paper indicates that established threat classes — command-and-control traffic, malicious domains, and script injection — continue to dominate encountered activity by volume, while a small but qualitatively distinct set of AI-workflow threats has emerged that is largely invisible to perimeter- and signature-based controls. Within the sample, the Salience Cyber User Edge platform mitigated 92.6% of events in real time, extended coverage into the AI and developer workflows where conventional tooling has no visibility and did so without a single user-facing interruption.

These findings, with the limitations of Section 7 noted, support a measured conclusion: securing the AI-enabled workplace does not require organizations to choose between enabling AI and defending against its associated risks. It requires a defensive posture that travels with the user, observes the AI channel directly, and combines signature-based precision with behavioral detection of the novel. Continued study across longer windows and more varied environments is warranted; the directional evidence presented here is, nonetheless, sufficiently clear to inform action.

References

[1] MITRE Corporation. (2024). MITRE ATT&CK: Adversary Groups — APT28 (G0007) and APT41 (G0096). Retrieved from https://attack.mitre.org/groups/

[2] Mandiant / Google Cloud. (2024). M-Trends 2024: Special Report. Retrieved from https://www.mandiant.com/m-trends

[3] Unit 42, Palo Alto Networks. (2022). Brute Ratel C4: The New Adversary Simulation Framework on the Block. Retrieved from https://unit42.paloaltonetworks.com/brute-ratel-c4-tool

[4] IBM Security. (2024). Cost of a Data Breach Report 2024. Retrieved from https://www.ibm.com/security/data-breach

[5] OWASP Foundation. (2021). OWASP Top 10:2021 — A03:2021 Injection. Retrieved from https://owasp.org/Top10/A03_2021-Injection

[6] Weichselbaum, L., Spagnuolo, M., Lekies, S., & Janc, A. (2016). CSP Is Dead, Long Live CSP! On the Insecurity of Whitelists and the Future of Content Security Policy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16).

Data sourced from the Salience Console (PostgreSQL/AWS RDS). Reporting period: 23 March – 1 April 2026. This paper presents a representative sample of operational telemetry and is intended to characterize directional trends and demonstrate platform behavior; it is not a population-level study. *· www.saliencecyber.ai*